# Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable? (3 marks)

Ans: Referring to the boxplot we created in the python notebook we can infer the following

1. Atemp and temp are highly correlated with each other, almost =1 so to avoid multicollinearity we drop the atemp column
2. Season is a good variable, can be used in the model as we see the demand increases during the fall and summer seasons.
3. There isn't much to gain from weekday as the median is almost same through all days so demand is not really changing much with days pf the week
4. Same can be concluded about holiday as there isn't a significant change in demand on the holidays

1. Why is it important to use **drop\_first=True** during dummy variable creation?

Ans: It is important to do so because the first column is not really needed to understand the data and it only ads a redundant dummy column to our dataset which in turn increases the number of columns. (2 mark)

1. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: Referring to our report from sweetviz, temperature has the highest correlation with dependent variable of 0.63

1. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Ans: By doing the residual analysis and also visualizing the same, the distribution of the residuals is normal and centered around mean 0 which is the main assumptions we have made

1. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

Ans: Temperature, season and month are affecting the demand most.

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks) 4

The linear regression algorithm is used in cases of variables that are linearly correlated with each other. Basically it fits a straight line through the variables and predictions are done on the basis of this line. Equation of the regression same as that of the line is y = mX + C

Where y is our variable that is to be predicted i.e dependent and x is our predictor. We choose x on the basis of a number of methods that are defined in the notebook and then evaluate these predictor variables.

1. Explain the Anscombe’s quartet in detail. (3 marks)

Ans: Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. Basically it demonstrates the importance of visualizing data and to show that summary statistics alone can be misleading.

1. What is Pearson’s R? (3 marks)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction. It is our measuring metric R squared that we use to evaluate the model.

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:  If not scaled, the feature with a higher value range starts dominating. The MinMaxscaler is a type of scaler that scales the minimum and maximum values to be 0 and 1 respectively. While the StandardScaler scales all values between min and max so that they fall within a range from min to max

(3 marks)

1. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When the value of R square is 1, mathematically the VIF will become infinity

(3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

(3 marks)